Deploying deep neural network with ONNX for efficient genome analysis with nanopore sequencing

Kishwar Shafin
Overview of genome analysis

Sample collection

Library Preparation

ACGTTACGTTATTCAGTTT
de novo assembly
Global effort
Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes.

Kishwar Shafin et al.
Nature Biotechnology, Accepted March 26, 2020
Shasta assembler

https://github.com/chanzuckerberg/shasta

Paolo Carnevali, CZI
Margin-Polish
H.E.L.E.N.

Multi-task learning with hard parameter sharing

https://github.com/kishwarshafin/helen
Base-level accuracy

![Bar graph showing base-level error rate (%)]

- **Shasta**:
  - HG00733: 0.98%
  - HG002: 1.06%
  - CHM13: 0.54%

- **Wtdbg2**:
  - HG00733: 1.18%
  - HG002: 1.22%
  - CHM13: 0.69%

- **Flye**:
  - HG00733: 1.64%
  - HG002: 1.85%
  - CHM13: 2.21%

- **Canu**:
  - HG00733: 1.40%
  - HG002: 1.33%
  - CHM13: 0.70%

**Average elapsed runtime (hour)**
- **Shasta**: 5.25 hr
- **Wtdbg2**: 62.53 hr
- **Flye**: 39.28 hr
- **Canu**: 80.00 hr

**Average cost ($)**
- **Shasta**: $70
- **Flye**: $142.03
- **Canu**: $695.21
Base-level accuracy improvement

Consensus sequence error rate (%)

- **Shasta**: Unpolished - 0.388%, Polished - 1.062%
- **Wtdbg2**: Unpolished - 0.473%, Polished - 1.217%
- **Canu**: Unpolished - 0.355%, Polished - 1.328%
- **Flye**: Unpolished - 0.356%, Polished - 1.854%

Legend:
- Red: Unpolished assembly
- Green: Polished with HELEN
Shasta run-time

Average elapsed runtime (hour)

- **Flye**: 62.53 hr, $695.21
- **Wtdbg2**: 39.28 hr, $142.03
- **Shasta**: 5.25 hr, $70
Run-time analysis

Average elapsed runtime (hour)

Average cost ($)
ONNX based HELEN CPU model deployment

MarginPolish images

ONNX session 1  ONNX session 2  ONNX session 3  ONNX session 4  ONNX session n

Predictions

Stitch

Polished sequence
Run-time improvement after introducing ONNX runtime

<table>
<thead>
<tr>
<th>Subset</th>
<th>Expected genome size</th>
<th>Previous release (Wall-clock time)</th>
<th>Current release (Wall-clock time)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human genome (HG00733)</td>
<td>3.2 Gb</td>
<td>40 hours</td>
<td>7 hours</td>
</tr>
<tr>
<td>Human genome (HG00733)</td>
<td>3.2 Gb</td>
<td>36 hours</td>
<td>6 hours</td>
</tr>
<tr>
<td>Human genome (CHM13)</td>
<td>3.2 Gb</td>
<td>42 hours</td>
<td>8 hours</td>
</tr>
<tr>
<td>Listeria monocytogenes (Microbial)</td>
<td>2.8 Mb</td>
<td>45 mins</td>
<td>6 mins</td>
</tr>
<tr>
<td>Bacillus subtilis (Microbial)</td>
<td>4.2 Mb</td>
<td>2 hours</td>
<td>12 mins</td>
</tr>
<tr>
<td>Salmonella enterica (Microbial)</td>
<td>5.1 Mb</td>
<td>3 hours</td>
<td>18 mins</td>
</tr>
<tr>
<td>Escherichia coli (Microbial)</td>
<td>4.6 Mb</td>
<td>2 hours</td>
<td>13 mins</td>
</tr>
</tbody>
</table>
Acknowledgements

UC SANTA CRUZ
David Haussler
Karen Miga
Ed Green
Sofie Salama
Hugh Olsen
Mark Akeson
Kristof Tigyi
Nicholas Maurer

Google Health
Pi-Chuan Chang
Andrew Carroll
Sidharth Goel
Maria Nattestad
Howard Yang

Adam Phillippy (NHGRI)
Sergey Koren (NHGRI)
Justin Zook (NIST)
Fritz Sedlazeck (Baylor)

Adam Novak
Glenn Hickey
Jordan Eizenga
Erik Garrison
Jean Monlong
Xian Chang

Daniel Garalde
Rosemary Dokos
Simon Mayes
Chris Seymour
Chris Wright
David Stoddart
Dan Turner
Vania Costa

Paolo Carnevali
Sidney Bell
Charlotte Weaver
Michael Barrientos
Ryan King
Bruce Martin
Phil Smoot
Cori Bargmann

Erich Jarvis
Chai Fungtammasan
Arang Rhie
+ Many More

Kelvin Liu
Duncan Kilburn

NIH National Human Genome Research Institute

CZ